

CLAIMS

What is claimed is:

1. A file content classification system comprising:
- 2 a digital ID generator;
- 3 an ID appearance database coupled to receive IDs from the ID
- 4 generator; and
- 5 a characteristic comparison routine identifying the file as having a
- 6 characteristic based on ID appearance in the appearance database.
2. The content classification system of claim 1 wherein said ID
- generator comprises a hashing algorithm.
3. The content classification system of claim 2 wherein said hashing
- algorithm is the MD5 hashing algorithm.
4. The content classification system of claim 1 wherein said ID
- appearance database tracks the frequency of appearance of a digital ID.
5. The content classification system of claim 1 further including a
- plurality of digital ID generators on different systems all coupled to and
- providing IDs to said ID appearance database.
6. The content classification system of claim 5 wherein said plurality
- of digital ID generators are coupled to said database via a combination of
- public and private networks.
7. The content classification system of claim 6 wherein said database
- is coupled to an intermediate server which is coupled to said plurality of

3 generators.

1 8. The content classification system of claim 6 wherein said  
2 intermediate server is a web server.

1 9. The content classification system of claim 1 wherein said  
2 characteristic comprises junk e-mail and said characteristic is defined by  
3 a frequency of appearance of a digital ID.

1 10. A method for identifying a characteristic of a data file, comprising:  
2 generating a digital identifier for the data file and forwarding the  
3 identifier to a processing system;  
4 determining whether the forwarded identifier matches a  
5 characteristic of other identifiers; and  
6 processing the email based on said step of determining.

1 11. The method of claim 10 wherein said step of generating comprises  
2 hashing at least a portion of the data file.

1 12. The method of claim 11 wherein said step of hashing comprises  
2 using the MD5 hash.

1 13. The method of claim 11 wherein said step of generating comprises  
2 hashing multiple portions of the data file.

1 14. The method of claim 10 wherein said data file is an email message  
2 and said step of determining comprises determining whether said email is  
3 spam.

1 15. The method of claim 10 wherein said step of determining identifies

2 said e-mail as spam by tracking the rate per unit time a digital ID is  
3 generated.

1 16. The method of claim 10 wherein said step of generating comprises  
2 generating IDs at a plurality of source systems all coupled via a network  
3 to at least one processing system performing the determining step.

1 17. The method of claim 16 wherein said step of processing comprises  
2 instructing said plurality of source systems to perform an action with the  
3 email based on said determining step.

1 18. A method of filtering an email message, comprising:  
2 processing the message to provide a digital identifier;  
3 comparing the digital identifier to a characteristic database of digital  
4 identifiers to determine whether the message has said characteristic; and  
5 processing the message based on said step of comparing.

1 19. The method of claim 18 wherein said step of processing occurs on  
2 at least one first system, and said step of comparing occurs on a second  
3 system.

1 20. The method of claim 19 wherein said step of processing occurs on  
2 a plurality of first systems.

1 21. The method of claim 19 wherein said at least one first system and  
2 second system are coupled by the Internet.

1 22. The method of claim 18 wherein said step of comparing comprises  
2 determining the frequency of a particular ID occurring in a time period,  
3 classifying said ID as having a characteristic, and comparing digital

4 identifiers to said classified IDs.

1 23. A file content classification system, comprising:  
2 a first system having a file to be classified;  
3 an file ID generator on the fist system;  
4 a database on a second system coupled to the ID generator to  
5 receive IDs generated by the ID generator;  
6 a comparison routine on the second system classifying the ID  
7 relative to the database as meeting or not meeting a characteristic.

1 24. The system of claim 23 including a plurality of first systems each  
2 including a respective file ID generator coupled to the database on the  
3 second system.

1 25. The system of claim 24 wherein the plurality of first systems is  
2 coupled to the second system via the Internet.

1 26. The system of claim 25 wherein the second system comprises a  
2 web server interface system and a database system, wherein the database  
3 system is isolated from the Internet by the web server system.

1 27. A content classification system for a first and second computer  
2 coupled by a network, comprising:  
3 a client agent file identifier generator on the first computer; and  
4 a server comparison agent and data-structure on the second  
5 computer receiving identifiers from the client agent and providing replies  
6 to the client agent;  
7 wherein the client agent processes the file based on replies from the  
8 server comparison agent.

1 31. The method of claim 29 wherein said step of characterizing  
2 comprises:  
3 tracking the frequency of the collection of a particular identifier;  
4 characterizing the data file based on said frequency;  
5 storing the characterization; and  
6 comparing collected identifiers to the known characterization.

add  
b<sup>i</sup>